PERCEPTUAL ADAPTIVE QUANTIZATION OF STEREOSCOPIC VIDEO CODING USING MPEG-2'S TEMPORAL SCALABILITY STRUCTURE

Belle L. Tseng & Dimitris Anastassiou

Center for Telecommunications Research
Department of Electrical Engineering Columbia University
530 West 120th Street CEPSR 801 New York, N.Y. 10027 USA
TEL: +1 (212) 854-6481 *FAX:* +1 (212) 316-9068 *EMAIL:* belle@ctr.columbia.edu

ABSTRACT

Significant perceptual improvements can be gained in the coding of stereoscopic videos, as is illustrated here by modifying the adaptive quantization of an MPEG-2 codec. Applying human visual properties to MPEG-2's temporal scalability framework, a perceptual adaptive quantization approach to stereoscopic video coding is presented. To optimize the perceived picture quality of the reconstructed stereo images, binocular visibility of stereo coding artifacts is investigated, including image fusion and visual masking. Four normalized indicators are introduced to account for 3D visual artifacts, and are incorporated to determine the quantization parameters. They are (1) prediction accuracy, (2) prediction correlation, (3) fusion indicator, and (4) texture intensity. Simulations indicate the importance of perceptual stereo coding, with improvements in overall stereo quality and reductions in binocular artifacts.

1. INTRODUCTION

Motivated by the idea of a combination 3D system and HDTV, we study the codec of 3DTV compatible with MPEG-2 standardization. Benefits and preferences for stereoscopic 3DTV have been shown in many applications, including medical imaging, remote handling, and quality control, where depth impression enhances the viewing experience and improves brilliance and fidelity. Because futuristic digital viewing displays vary in stereoscopic capabilities, bit rate allocation and quantization control impose important restrictions during the encoding stage. It is for this goal that we propose a perceptual adaptive quantization for a 3DTV codec compatible with MPEG-2 standardization.

Basic transmission of a stereoscopic video sequence consists of a left and right channel, each carrying images captured from the corresponding views.[1] With the objective of compression and transmission on one bandwidth-limited channel, efficient coding of the two video signals is achieved by exploiting their interrelationships. Within the MPEG-2 standardization, it have been shown that transmission of a stereoscopic sequence is possible by utilizing the high profile double layer structure of temporal scalable coding, as discussed with more detail in Section 2.

Adopting the temporal scalable feature of MPEG-2 for stereoscopic video processing, our goal is to optimize the perceived picture quality of the reconstructed images. The scheme is to improve the current conventional adaptive quantization control, so as to incorporate the binocular perceptual factors involved in stereo viewing. In Section 3, current bit rate and quantization controls based on MPEG-2's Test Model, which is mainly dependent on 2D spatial activity indicators, are reviewed. In Section 4, a survey of human perceptual factors from the viewpoint of stereo vision is examined, including binocular fusion and stereo artifact masking.

Since our human visual perception of a stereo pair depends on two correlated images, we adapt and quantize the DCT coefficients accordingly. Four interdependent classes of parameters are known to be crucial in binocular perception, and are utilized in determining the quantization parameters. The strategy and details of which are discussed in Section 5.

2. MPEG-2'S TEMPORAL SCALABILITY

The primary objective of the temporal scalability feature of MPEG-2 is to accommodate future sophisticated high temporal resolution systems. Temporal scalability, similar to the other scalability profiles, is built upon the main framework of MPEG-2.[2] The base layer codec performs equivalent operations as the non-scalable video coding. In addition, the codec for the enhancement layer is similar to that for the lower layer. The difference being that the motion compensated predictions are now with respect to frames from either the base or the enhancement layer.

For stereoscopic video transmission, the temporal scalability feature of MPEG-2 offers two interdependent layers of coding[3], as illustrated in Figure 1. The left stereo sequence is coded on the lower layer and provides the basic non-stereoscopic signal. The right stereo bitstream is then transmitted on the enhancement layer and when combined with the left view results in the full stereoscopic video. The coding specification allows each right image r_i on the enhancement layer to be disparity-estimated from the corresponding left image l_i of the base layer, or motion-estimated from the previously

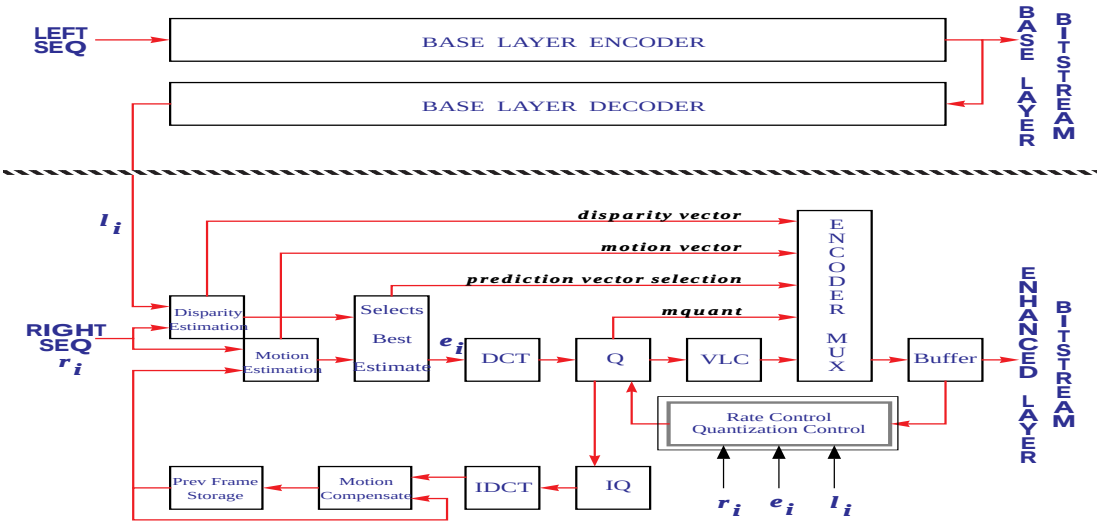

Figure 1: MPEG-2's Two-Layer Temporal Scalable Coding

encoded right image. In this manner, the enhancement layer selects the best prediction combination of spatial (left → right) and temporal (right → right) information.

In the temporal scalable syntax, the disparity estimates are transmitted as prediction vectors which are embedded in the basic motion compensated framework of MPEG-2. As a result, any MPEG-2 codec equipped with temporal scalable coding can universally encode and decode stereoscopic sequences without requiring any further implementation changes; except for modifying the disparity vector search range in the encoder. In application to stereo signals, temporal scalable video coding offers MPEG-2 compatibility, easy and direct implementation convenience, graceful stereo image degradation, and high SNR reconstructions.[3]

After selecting the best prediction estimate, the prediction error e_i is calculated with respect to the original right signal. Following, similar to the main profile of MPEG-2, three additional stages are performed to compress the prediction error: computation of the discrete cosine transform (DCT) coefficients; quantization of the DCT coefficients; and conversion into variable length codes. As shown in Figure 1, the quantization task is dependent on a quantization parameter output by the Rate Control block module, marked with a double box. In our modification of the block module, additional image data, l_i, r_i, and e_i, are examined before determining the rate control parameters.

3. BIT RATE AND QUANTIZATION CONTROLS

The bit rate of an encoded video bitstream largely influences the quality of the reconstructed sequence. An elaborate 2D-perceptual adaptive quantization strategy based on motion compensated video coding framework of MPEG has been developed in [4] to improve the perceived image sequence. The bit rate controller used by most encoding systems is based on MPEG-2's Test Model 4,[5] where one quantization parameter ultimately dictates the quantization of the DCT coefficients in a macroblock. In this section, a review of this conventional controller is presented by analyzing its three major components.

In step 1, the target number of bits for each frame is estimated before encoding of its picture, which is dependent on its picture type. Initially, global complexity estimation for each picture type is determined from previously encoded frames. Following, target bit allocation for each picture type is calculated based on a complexity measure, and as expected is also proportional to the number of available remaining bits. After encoding of another picture, all parameters are updated.

In step 2, the reference quantization parameter (Q_{ref}) is determined for each macroblock in accordance with the available bits allocated for that picture and the fullness of a "virtual" buffer. Finally in the adaptive quantization component of step 3, the value of the quantization parameter ($mquant$) is calculated for each macroblock, with respect to its reference Q_{ref}, in the following series of parameter computation:

$$\text{variance of subblock } i = var(subblock\ i) \text{ where } i \in [1,4]$$

$$\text{2D spatial activity } act = 1 + \min var(subblock\ i)$$

$$avg\_act = \text{average } act \text{ values over the last encoded frame}$$

$$\text{normalized spatial activity } N\_act = \frac{2*act+avg\_act}{act+2*avg\_act}$$

$$\text{quantization parameter } mquant = Q_{ref} * N\_act.$$

Quantization is thus predominantly controlled by the lowest textured subblock, which is visually least tolerable to 2D artifacts. Thus by observation, the maximum value of the quantization parameter $mquant = 2 * Q_{ref}$ with respect to the calculated reference. Similarly, the minimum achievable value for $mquant = \frac{1}{2} * Q_{ref}$.

4. HUMAN PERCEPTUAL FACTORS

The human visual system converts two slightly displaced viewpoint images into one fused 3D perception. Not only does stereoscopic vision provide us with depth discrimination, but the single perceived view merges and provides information from two monocular images, although highly overlapping. The physiological mechanisms involved in binocular fusion are still unclear, but some important results have been demonstrated as will be briefly reviewed in this section.

When two images of the world are received by the left and right eyes, automatic pairing between areas of the two images is performed in the human visual cortex mainly based on edged features.[6] In parallel, when two images are captured by a left and right camera, pairing between similar regions of the left and right images needs to be extracted, also known as the *correspondence problem*. It follows that once the correspondence problem is solved, the depth has also been determined.

Once correspondence is achieved between coupled regions, two complementary theories have been suggested in the combining process.[7] One is the *Suppression Theory* where one prevailing region dominates the visual perception for that space while the other corresponding region is suppressed. Suppression Theory, which gives way to *binocular rivalry*, usually results in an alternating cycle of dominance and suppression between the two distinct regions. The other idea is the *Fusion Theory* in which the two coupled regions are fused together in the visual field as a single representative. Fusion Theory, which is the result of achieving *binocular cooperation*, unfortunately does not describe how the individual left and right counterparts are merged.

The two theories, although mutually exclusive, have been demonstrated to be both applicable to the human visual system, thus vision experts have been suggesting a combination theory as compromise. Suppression Theory dominates when the two corresponded areas tend to be excessively inconsistent, while Fusion Theory prevails when the matched regions are more or less similar. *Binocular fusion*, whose terminology refers to the fusion mechanism undertaken by the human cortical neurons designed to merge two correspondingly similar images, is still an open research topic. One suggested method is *binocular masking* where the sharper (more contrasting) region dominates the visual field. Thus artifact visibility is reduced when one appropriately degraded image is fused with a complete image, compensating each other to yield a perceptually acceptable stereo picture.[8]

5. PERCEPTUAL ADAPTIVE QUANTIZATION

5.1. Overview

Adopting the MPEG-2 coding framework and utilizing the binocular masking property, the objective is to perceptually quantize stereoscopic signals so that compression artifacts are visually concealed. Since the conventional bit rate controller adapts the quantization parameter only to 2D activity indicators, a modification on the current adaptive quantization step is proposed. Without loss of generality and as illustrated in Figure 1, the left stereo sequence is encoded in the base layer using conventional adaptive quantization, and the right sequence in the enhancement layer with the modified version. It is therefore the enhancement quantization controller with left and right stereo information availability that we are concentrating on. In addition, similar to the conventional quantization scheme, the proposed strategy is one-pass and causal, where calculations are based on performances of previously encoded data.

After determining the reference quantization parameter Q_{ref} for a macroblock, the lower and upper bounds are approximated as $Q_{ref\_lo} = \frac{1}{2}Q_{ref}$ and $Q_{ref\_hi} = \frac{3}{2}Q_{ref}$ respectively. Next four classes of normalized perceptual indicators are determined: (1) Prediction Accuracy, (2) Prediction Correlation, (3) Fusion Indicator, and (4) Texture Intensity. Depending on these four normalized values, the quantization parameter $mquant$ is assigned based on 3D visual masking properties.

5.2. Perceptual Indicator Classification

When two corresponding regions of a stereo pair are fused by the human visual system, the perceived image depends on whether Suppression Theory or Fusion Theory dominates. Thus if matched correspondence pairs are excessively dissimilar, as in the case of obstructed objects appearing in only one view, then we need never consider the binocular fusion mechanism in this situation. However for correspondence pairs in which binocular fusion is achieved, then binocular masking of certain stereo coding artifacts may occur. We have formulated 4 classes of macroblock-unit perceptual indicators to account for these artifacts. The indicators are based on statistical activities of the original right macroblock r_i, the residual error macroblock e_i in encoding the right image, and the decoded left macroblock l_i which is in correspondence with the right input at hand. For each macroblock, index $i \in [1, 4]$ represents each of the four 8x8 luminance subblocks. A set of adjustable parameters is utilized, where $T1$, $T2$, $T3$ & K represent integral thresholds, and $C1$, $C2$, $C3$ & $C4$ denote real constants.

The first normalized indicator $p1$ measures **prediction accuracy** in encoding the right input. For nearly perfect predictions, $p1 = 1$, and for visually-intolerable wrong predictions, $p1 = 0$. Prediction accuracy is measured as:

$$p1 = 1 - \mathsf{TRUNC}(C1 * \frac{\mathsf{AVE}\ mse(e_i)}{\mathsf{AVE}\ var(r_i)}).$$

Let $\mathsf{TRUNC}(x)$ mean truncating the real value x, so that if $x < 0$ then $x = 0$, and similarly if $x > 1.0$ then $x = 1.0$. Thus $\mathsf{TRUNC}(x) \in [0, 1]$. mse denotes mean square error computation, and var defines the variance calculation. AVE, MAX, and MIN are determined over subblock index $i \in [1, 4]$.

The second indicator $p2$ measures **prediction correlation** between the original right signal and the compensated prediction. If the prediction is well correlated with the original thus visually negligible, then $p2 = 1$; on the other hand, if the prediction resembles a visually disturbing blocky version of the original, then $p2 = 0$. Initially, extremely high correlations are detected by the following:

$$\textbf{If} \quad \mathsf{MAX}\ mse(e_i) < T1 \quad \textbf{or} \quad \mathsf{MAX}\ chr\_mse(e_i) < T2 \quad \textbf{or} \quad \mathsf{MAX}\ var(e_i) < T3,$$

then $p2 = 1$. Correspondingly for truly uncorrelated predictions with biased distributions of subblock statistics like in a blocky reconstruction:

$$\textbf{If} \quad \frac{\mathsf{MIN}\ f(i)}{\mathsf{AVE}\ f(i)} < C2 \quad \textbf{or} \quad \frac{\mathsf{MAX}\ f(i)}{\mathsf{AVE}\ f(i)} > \frac{1}{C2},$$

then $p2 = 0$, where the notation $f(i)$ refers to $mse(e_i)$ and $chr\_mse(e_i)$ and $var(e_i)$. chr_mse denotes mse computation of the chrominance components. To detect the non-extreme range of $p2 \in (0, 1)$,

$$p2 = 1 - \sum \frac{1}{4} \mathsf{TRUNC}(C3 * \frac{var(e_i)}{var(r_i)}).$$

The third classification f is the **fusion indicator** to detect whether Fusion Theory is dominant in this macroblock. High fusion, $f = 1$, is associated with high similarity between corresponding left and right regions. On the contrary, no fusion, $f = 0$, occurs when Suppression Theory prevails which usually suggests an occluded area or an inaccurate disparity estimate. The normalized fusion indicator is measured as:

$$f = 1 - C4 * \mathsf{TRUNC}(\frac{\mathsf{AVE}\ mse(l_i - r_i)}{\mathsf{AVE}\ var(r_i)}) \ - \ (1 - C4) * \mathsf{TRUNC}(\frac{\mathsf{AVE}\ chr\_mse(l_i - r_i)}{\mathsf{AVE}\ var(r_i)}),$$

where the proportional constant $C4$ weights the luminance and chrominance information.

Finally, the fourth normalized indicator t incorporates the 2D spatial activity and estimates the **texture intensity**. For highly structured textured areas, $t = 1$, whereas in no or low textured areas, $t = 0$. Assign $t = 0$ to a flat luminance macroblock, and $t = \frac{1}{2}$ for an area corresponding to the global average texture, avg_act, and $t = 1$ for a region whose texture activity is greater than K times the global average. The following monotonic function results,

$$t = \mathsf{TRUNC}(\frac{(K-1) * act}{(K-2) * act + K * avg\_act}),$$

where act and avg_act are defined in Section 3.

5.3. Determining the Quantization Parameter

After obtaining the normalized indicators $p1$, $p2$, f, and t, the task is to determine the quantization parameter $mquant \in [Q_{ref\_lo}, Q_{ref\_hi}]$ based on human perceptual tolerance to stereo coding artifacts. The binocular coding artifacts which are accounted for in our presentation include blockiness, inaccuracy, and blurriness.

In viewing stereo artifacts, we have noticed that blocky reconstructions in one image are visually annoying even when the other stereo image is very well coded, thus no binocular masking reductions of such artifacts are possible. Similarly when an inaccurate reconstruction is viewed with a well coded counterpart, the overall stereo perception of the pair is visually undesirable, and alternating suppression may occur on the artifact region. Thus, if prediction results in a poor ($p1 \approx 0$) or blocky ($p2 \approx 0$) reconstruction, then quantization must be refined ($mquant \rightsquigarrow Q_{ref\_lo}$). On the other hand, if prediction is highly accurate ($p1 \approx 1$) and well correlated ($p2 \approx 1$) with respect to the original, then quantization can be coarse ($mquant \rightsquigarrow Q_{ref\_hi}$). Consequently, an average prediction $\frac{p1+p2}{2}$ is calculated to represent the overall perceptual acceptability.

When two corresponding regions of the left and right images have a high probability of binocular fusion due to the highly similar characteristics of the matched areas, then fusion theory states that our human visual system perceive the paired images as one. It follows that if the original right macroblock achieves high fusion with a corresponding decoded left region, then it is usually the case that the better coded region of the pair dominates the visual field, thus the right macroblock may be directly predicted (spatially) from its corresponding left image with little or no additional correction. By deduction, if the *fusion indicator* is high ($f \approx 1$), then quantization can be coarse ($mquant \rightsquigarrow Q_{ref\_hi}$). However, if the *fusion indicator* is poor ($f \approx 0$), then quantization should be refined ($mquant \rightsquigarrow Q_{ref\_lo}$).

Another common 2D coding artifact that occurs frequently in low bit rate transmission is blurriness. However in stereo viewing experiments[9][10], a stereo image pair, where one of the images is blurred and the other close to the original, is perceived as a sharp 3D picture, maintaining the local contrast information from the well coded image. It then follows that as long as fusion of corresponding left and right regions is possible ($f \approx 1$), perceptual acceptability is increased and encoding may be relaxed ($\uparrow mquant$). As a result, a high fusion indicator means the corresponding well-encoded left region have a higher probability of covering up possible artifact in the right macroblock.

In regard to the binocular fusability of the left and right regions, if texture is highly structured ($t \approx 1$), then the more accurate the fusability indicator f, the better the correspondence match, and the superior the disparity estimated prediction. Furthermore, the texture information derived from 2D spatial activity act is used to tune the final quantization parameter. To wrap up, we have compacted the following measure to derive the final $mquant$.

$$w = \frac{p1}{2} + \frac{p2}{2} + \frac{f}{4} + \frac{t}{8} - \frac{3}{16},$$

where the final constant $\frac{3}{16}$ offsets the parameter w, so that $[p1, p2, f, t] = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}] \quad \Rightarrow \quad w = \frac{1}{2}$. Finally,

$$mquant = Q_{ref\_lo} + w * (Q_{ref\_hi} - Q_{ref\_lo}).$$

6. SIMULATIONS AND CONCLUSION

Simulations are performed using Columbia's MPEG-2 software package, which includes the temporal scalability feature. In addition, the existing formats of MPEG-2 have been extended for stereoscopic video coding[3], with modifications on the enhancement adaptive quantization module according to the presented work. In our simulations, unequal bandwidths are allocated to the left base-layer and right enhancement-layer stereo coding, where the right sequence is transmitted at the lower bit rate for possible applicability of stereo artifact masking. Quantitative performance measures are based on the luminance signal to noise ratio (SNR) of the decoded reconstructions with respect to the original images.

Like most quantization controller, parametric training is dependent on the stereo sequence and the total transmission bandwidth. Values for the threshold parameters are initially assigned but updated after encoding of every frame, with starting values of $T1 = 150$, $T2 = 50$, $T3 = 10$, $K = 4$, $C1 = 1.0$, $C2 = 0.4$, $C3 = 0.2$, & $C4 = 1.0$. Parameter re-estimation is based on balancing the four normalized indicators so that $[p1, p2, f, t] = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}] \quad \Rightarrow \quad w = \frac{1}{2} \quad \Rightarrow \quad mquant = Q_{ref}$.

In our experimental simulations, 4 stereoscopic video sequences courtesy of CCETT[11], *Aqua, Discussion, Manege,* and *Train*, are used in the perceptual performance evaluations. Each stereoscopic signal consists of a left and right interlaced sequence, at European CCIR 601 resolution (720x576 at 25 frames/sec) with 422 chroma format. The right sequence is coded at approximately half the bit rate allocated for the left sequence as suggested by [12]. Table 1 itemizes the *averaged* performances of the left and right stereo sequences for our perceptual experiments with the simulcast approach and the proposed method. Figure 2 graphically illustrates these improvements over the simulcast results for two of the sequences.

As compared to the conventional quantization scheme, the presented formulation results in perceptual reduction of blocky artifacts and elimination of inaccurate macroblock reconstructions. In addition, the overall sharpness contrast improves by compensating areas that have low/no fusable counterpart with areas that do, thus taking advantage of the binocular masking effect. Besides, the modified quantization also demonstrates quantitative improvements of about $0.1 \sim 0.3$dB over the original quantization of the same temporal scalability codec at each transmission bit rate.

Utilizing 3D perceptual masking properties, the proposed modified adaptive quantization illustrates the significance of visual stereo information verses using 2D spatial activity information alone. As new advances in binocular vision develop, the quantization algorithm may be modified to conform. Only a small aspect of the human stereo factors have been incorporated in the presented work, however it demonstrates the importance and necessity for perceptual coding in stereoscopic video coding. In conclusion, an MPEG-2 compatible framework for future 3D perceptual studies is provided for further research.

7. REFERENCES

[1] M. Ziegler et al, "Digital stereoscopic television - state of the European project DISTIMA," in *4th European Workshop on Three-Dimensional Television*, (Rome, Italy), Oct. 1993.

[2] "MPEG Committee Draft. ISO/IEC 13818-2: Generic coding of moving pictures and associated audio," Mar. 1994.

[3] B. L. Tseng and D. Anastassiou, "Compatible video coding of stereoscopic sequences using MPEG-2's scalability and interlaced structure," in *International Workshop on HDTV '94*, (Torino, Italy), Oct. 1994.

[4] A. Puri and R. Aravind, "Motion-compensated video coding with adaptive perceptual quantization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, Dec. 1991.

[5] "MPEG Proposal for Test Model 4, Draft Revision 1. coded representation of picture and audio information," 1993.

[6] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*. New York: Elsevier Science Publishing Co., Inc., third ed., 1991.

[7] "Binocular vision," in *Vision and Visual Dysfunction* (D. Regan, ed.), vol. 9, CRC Press, 1991.

[8] L. Dinstein, M. Kim, J. Tselgov, and A. Henik, "Compression of stereo images and the evaluation of its effects on 3-D perception," *SPIE Vol. 1153 Applications of Digital Image Processing XII*, pp. 522–530, 1989.

[9] S. Pastoor, M. Wopking, J. Fournier, and T. Alpert, "Digital stereoscopic imaging and applications (DISTIMA): Human factors data," *RACE II - R2045 - DISTIMA Deliverable 26*, Feb. 1994.

[10] S. Pastoor, "3D-television: A survey of recent research results on subjective requirements," *Signal Processing: Image Communication*, vol. 4, pp. 21–32, 1991.

[11] "Stereoscopic Test Sequences AQUA, DISCUSSION, MANEGE, and TRAIN (CCIR 601 format) shot within DISTIMA."

[12] B. Choquet, F. Chassaing, J. Fournier, D. Pele, A. Poussier, and H. Sanson, "3D TV studies at CCETT," in *Proceedings TAO First International Symposium*, (Tokyo, Japan), Dec. 1993.

	Simulcast B=2Mb/s E=2Mb/s T=4Mb/s	*Temporal* B=2Mb/s E=2Mb/s T=4Mb/s	*Temporal* B=3Mb/s E=2Mb/s T=5Mb/s	*Simulcast* B=3Mb/s E=3Mb/s T=6Mb/s	*Temporal* B=4Mb/s E=2Mb/s T=6Mb/s	*Temporal* B=4Mb/s E=3Mb/s T=7Mb/s	*Simulcast* B=4Mb/s E=4Mb/s T=8Mb/s	*Temporal* B=5Mb/s E=3Mb/s T=8Mb/s
AQUA								
Left View SNR	32.50	32.50	34.35	34.35	35.72	35.72	35.72	36.64
Right View SNR	30.48	31.69	32.04	32.25	32.27	33.53	33.56	33.68
L-R Average SNR	31.49	32.10 (+0.61)	33.20 (+1.71)	33.30 (+1.81)	34.00 (+2.51)	34.63 (+3.14)	34.64 (+3.15)	35.16 (+3.67)
DISCUSSION								
Left View SNR	29.08	29.08	30.88	30.88	32.16	32.16	32.16	33.22
Right View SNR	29.04	29.60	29.87	30.77	30.01	31.54	32.02	31.88
L-R Average SNR	29.06	29.34 (+0.28)	30.38 (+1.32)	30.83 (+1.77)	31.08 (+2.02)	31.85 (+2.79)	32.09 (+3.03)	32.55 (+3.49)
TRAIN								
Left View SNR	31.31	31.31	33.88	33.88	35.36	35.36	35.36	36.43
Right View SNR	31.37	33.78	34.53	33.83	34.98	36.15	35.26	36.40
L-R Average SNR	31.34	32.55 (+1.21)	34.20 (+2.86)	33.85 (+2.51)	35.17 (+3.83)	35.76 (+4.42)	35.31 (+3.97)	36.42 (+5.08)
MANEGE ($L \rightarrow R$)								
Left View SNR	26.68	26.68	28.57	28.57	30.10	30.10	30.10	31.33
Right View SNR	28.25	29.40	29.69	30.40	29.91	31.71	31.84	31.82
L-R Average SNR	27.46	28.04 (+0.58)	29.13 (+1.67)	29.49 (+2.03)	30.00 (+2.54)	30.91 (+3.45)	30.97 (+3.51)	31.58 (+4.12)
MANEGE ($R \rightarrow L$)								
Right View SNR	28.25	28.25	30.40	30.40	31.84	31.84	31.84	32.99
Left View SNR	26.68	27.65	27.95	28.57	28.13	29.98	30.10	30.13
L-R Average SNR	27.46	27.95 (+0.49)	29.18 (+1.72)	29.49 (+2.03)	29.99 (+2.53)	30.91 (+3.45)	30.97 (+3.51)	31.56 (+4.10)

Table 1: Luminance SNR Performances of the Conventional Simulcast Approach (*Simulcast*) and the Proposed Perceptual Adaptive Quantized Temporal Scalability Scheme (*Temporal*), where B = Base Layer Bandwidth, E = Enhancement Layer Bandwidth, and T = Total Transmission Bandwidth. In each simulation, the Left Stereoscopic Sequence is transmitted on the Base Layer and the Right Sequence on the Enhancement Layer, except for the last experiment *Manege* ($R \rightarrow L$).

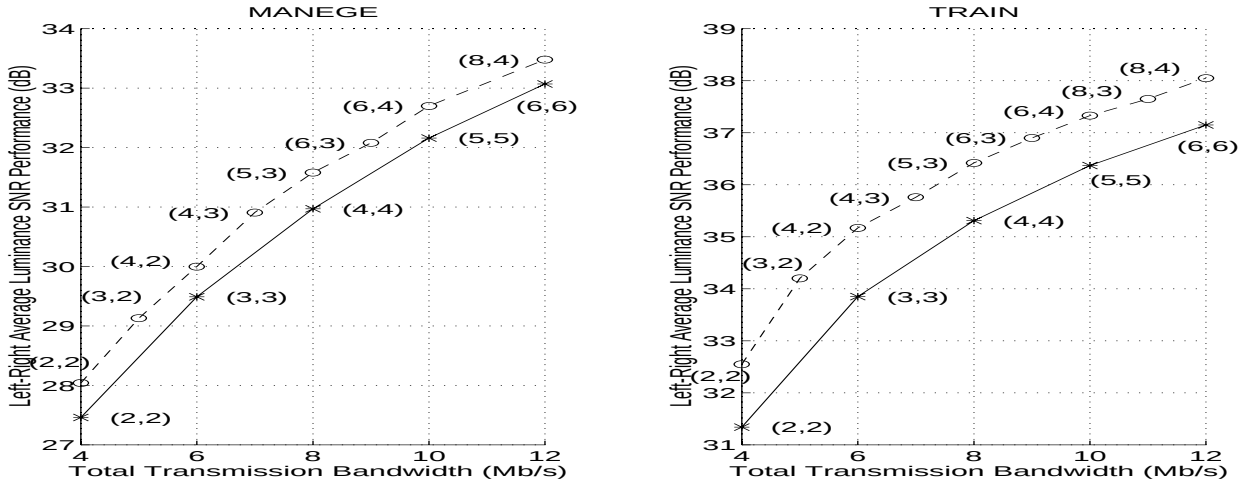

Figure 2: Left and Right Averaged Luminance SNR Performance Comparison between the Conventional Simulcast Approach (*solid line*) and the Proposed Perceptual Adaptive Quantization Scheme using the Temporal Scalability Structure (*dashed line*). Performances are plotted as a function of Total Transmission Bandwidth with individual stereoscopic sequence bit rate allocations denoted by (L, R) where L is the Left and R is the Right.